My Roll No. Is : DS5B-2137

Creating Linear Regression Model Using PySpark

Install PySpark

In [ ]:
```
pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/publ
ic/simple/
Collecting pyspark
  Downloading pyspark-3.2.1.tar.gz (281.4 MB)
     |████████████████████████████████| 281.4 MB 34 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 46.5 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.1-py2.py3-none-any.whl size=281853642
sha256=99657e37a6edb52a83d4b4e280e11c4e26947120a4b2dc8192749712f371238e
  Stored in directory: /root/.cache/pip/wheels/9f/f5/07/7cd8017084dce4e93e84e92efd1e1d53
34db05f2e83bcef74f
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.3 pyspark-3.2.1
```

Import Library and Creating Session

In [ ]:
```
from pyspark.sql import SparkSession
```

In [ ]:
```
session = SparkSession.builder.appName("exam1").master("local").getOrCreate()
```

Read Dataset

In [ ]:
```
data = session.read.csv("Big Mart Sale.csv", header = True, inferSchema=True)
```

To print top 10 raw in dataset

In [ ]:
```
data.show(10)
```

```
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|           Item_Type|Item_
MRP|Outlet_Identifier|Outlet_Establishment_Year|Outlet_Size|Outlet_Location_Type|      O
utlet_Type|Item_Outlet_Sales|
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
|          FDA15|        9.3|         Low Fat|    0.016047301|               Dairy|249.8
092|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|         3735.138|
|          DRC01|       5.92|         Regular|    0.019278216|         Soft Drinks| 48.2
692|           OUT018|                     2009|     Medium|              Tier 3|Superma
rket Type2|         443.4228|
|          FDN15|       17.5|         Low Fat|    0.016760075|                Meat| 141.
618|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|          2097.27|
|          FDX07|       19.2|         Regular|            0.0|Fruits and Vegeta...| 182.
095|           OUT010|                     1998|       null|              Tier 3|    Gro
cery Store|           732.38|
|          NCD19|       8.93|         Low Fat|            0.0|           Household| 53.8
614|           OUT013|                     1987|       High|              Tier 3|Superma
rket Type1|         994.7052|
|          FDP36|     10.395|         Regular|            0.0|        Baking Goods| 51.4
008|           OUT018|                     2009|     Medium|              Tier 3|Superma
rket Type2|         556.6088|
|          FDO10|      13.65|         Regular|    0.012741089|         Snack Foods| 57.6
588|           OUT013|                     1987|       High|              Tier 3|Superma
rket Type1|         343.5528|
|          FDP10|       null|         Low Fat|    0.127469857|         Snack Foods|107.7
622|           OUT027|                     1985|     Medium|              Tier 3|Superma
rket Type3|        4022.7636|
|          FDH17|       16.2|         Regular|    0.016687114|        Frozen Foods| 96.9
726|           OUT045|                     2002|       null|              Tier 2|Superma
rket Type1|        1076.5986|
|          FDU28|       19.2|         Regular|     0.09444959|        Frozen Foods|187.8
214|           OUT017|                     2007|       null|              Tier 2|Superma
rket Type1|         4710.535|
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
only showing top 10 rows
```

Check Null Values in columns

In [ ]:

```python
from pyspark.sql.functions import isnan, when, count, col
```

In [ ]:

```python
data.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c) for c in data.columns])
```

```
+---------------+-----------+----------------+---------------+---------+--------+-------
----------+-------------------------+-----------+--------------------+-----------+------
-----------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|Item_Type|Item_MRP|Outlet_
Identifier|Outlet_Establishment_Year|Outlet_Size|Outlet_Location_Type|Outlet_Type|Item_O
utlet_Sales|
+---------------+-----------+----------------+---------------+---------+--------+-------
----------+-------------------------+-----------+--------------------+-----------+------
-----------+
|              0|       1463|               0|              0|        0|       0|       
         0|                        0|       2410|                   0|          0|      
          0|
+---------------+-----------+----------------+---------------+---------+--------+-------
----------+-------------------------+-----------+--------------------+-----------+------
-----------+
```

In [ ]:

```python
import pyspark.sql.functions as func
```

In [ ]:

```python
data.agg(func.percentile_approx("Item_Weight", 0.5).alias("mean")).show()
```

```
+-----+
|mean|
+-----+
|12.6|
+-----+
```

Fill Null Values

First we replace 12.6 in place of Null values in Item_weight column because it is mean in this column

```
In [ ]: data = data.na.fill(value=12.6,subset=["Item_Weight"])
```

Second we return Medium in place of Null values in Outlet_Size Column Because Medium is the median in Outlet_Size Column

```
In [ ]: data = data.na.fill(value="Medium",subset=["Outlet_Size"])
```

```
In [ ]: data.show()
```

```
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|           Item_Type|Item_
MRP|Outlet_Identifier|Outlet_Establishment_Year|Outlet_Size|Outlet_Location_Type|      O
utlet_Type|Item_Outlet_Sales|
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
|          FDA15|        9.3|         Low Fat|    0.016047301|               Dairy|249.8
092|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|         3735.138|
|          DRC01|       5.92|         Regular|    0.019278216|         Soft Drinks| 48.2
692|           OUT018|                     2009|     Medium|              Tier 3|Superma
rket Type2|         443.4228|
|          FDN15|       17.5|         Low Fat|    0.016760075|                Meat| 141.
618|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|          2097.27|
|          FDX07|       19.2|         Regular|            0.0|Fruits and Vegeta...| 182.
095|           OUT010|                     1998|     Medium|              Tier 3|    Gro
cery Store|           732.38|
|          NCD19|       8.93|         Low Fat|            0.0|           Household| 53.8
614|           OUT013|                     1987|       High|              Tier 3|Superma
rket Type1|         994.7052|
|          FDP36|     10.395|         Regular|            0.0|        Baking Goods| 51.4
008|           OUT018|                     2009|     Medium|              Tier 3|Superma
rket Type2|         556.6088|
|          FDO10|      13.65|         Regular|    0.012741089|         Snack Foods| 57.6
588|           OUT013|                     1987|       High|              Tier 3|Superma
rket Type1|         343.5528|
|          FDP10|       12.6|         Low Fat|    0.127469857|         Snack Foods|107.7
622|           OUT027|                     1985|     Medium|              Tier 3|Superma
rket Type3|        4022.7636|
|          FDH17|       16.2|         Regular|    0.016687114|        Frozen Foods| 96.9
726|           OUT045|                     2002|     Medium|              Tier 2|Superma
rket Type1|        1076.5986|
|          FDU28|       19.2|         Regular|     0.09444959|        Frozen Foods|187.8
214|           OUT017|                     2007|     Medium|              Tier 2|Superma
rket Type1|         4710.535|
|          FDY07|       11.8|         Low Fat|            0.0|Fruits and Vegeta...| 45.5
402|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|        1516.0266|
|          FDA03|       18.5|         Regular|    0.045463773|               Dairy|144.1
102|           OUT046|                     1997|      Small|              Tier 1|Superma
rket Type1|         2187.153|
|          FDX32|       15.1|         Regular|      0.1000135|Fruits and Vegeta...|145.4
786|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|        1589.2646|
|          FDS46|       17.6|         Regular|    0.047257328|         Snack Foods|119.6
782|           OUT046|                     1997|      Small|              Tier 1|Superma
```

```
rket Type1|        2145.2076|
|          FDF32|      16.35|         Low Fat|      0.0680243|Fruits and Vegeta...|196.4
426|           OUT013|                     1987|       High|              Tier 3|Superma
rket Type1|         1977.426|
|          FDP49|        9.0|         Regular|    0.069088961|           Breakfast| 56.3
614|           OUT046|                     1997|      Small|              Tier 1|Superma
rket Type1|        1547.3192|
|          NCB42|       11.8|         Low Fat|    0.008596051|  Health and Hygiene|115.3
492|           OUT018|                     2009|     Medium|              Tier 3|Superma
rket Type2|        1621.8888|
|          FDP49|        9.0|         Regular|    0.069196376|           Breakfast| 54.3
614|           OUT049|                     1999|     Medium|              Tier 1|Superma
rket Type1|         718.3982|
|          DRI11|       12.6|         Low Fat|    0.034237682|         Hard Drinks|113.2
834|           OUT027|                     1985|     Medium|              Tier 3|Superma
rket Type3|         2303.668|
|          FDU02|      13.35|         Low Fat|     0.10249212|               Dairy|230.5
352|           OUT035|                     2004|      Small|              Tier 2|Superma
rket Type1|        2748.4224|
+---------------+-----------+----------------+---------------+--------------------+-----
---+-----------------+-------------------------+-----------+--------------------+-------
----------+-----------------+
only showing top 20 rows
```

A Simple Exploratory Of Dataset

To print all columns name

In [ ]:
```
data.columns
```

Out[ ]:
```
['Item_Identifier',
 'Item_Weight',
 'Item_Fat_Content',
 'Item_Visibility',
 'Item_Type',
 'Item_MRP',
 'Outlet_Identifier',
 'Outlet_Establishment_Year',
 'Outlet_Size',
 'Outlet_Location_Type',
 'Outlet_Type',
 'Item_Outlet_Sales']
```

To count total numbers of raws in dataset

In [ ]:
```
data.count()
```

Out[ ]:
```
8523
```

To print the schema of dataset, Spark schema is the structure of the DataFrame or Dataset, which is a collection of StructField that define the column name(String), column type (DataType), nullable column (Boolean) and metadata (MetaData)

In [ ]:
```
data.printSchema()
```

```
root
 |-- Item_Identifier: string (nullable = true)
 |-- Item_Weight: double (nullable = false)
 |-- Item_Fat_Content: string (nullable = true)
 |-- Item_Visibility: double (nullable = true)
 |-- Item_Type: string (nullable = true)
 |-- Item_MRP: double (nullable = true)
 |-- Outlet_Identifier: string (nullable = true)
 |-- Outlet_Establishment_Year: integer (nullable = true)
 |-- Outlet_Size: string (nullable = false)
 |-- Outlet_Location_Type: string (nullable = true)
 |-- Outlet_Type: string (nullable = true)
 |-- Item_Outlet_Sales: double (nullable = true)
```

To know data type of each columns

In [ ]:

```
data.dtypes
```

Out[ ]:

```
[('Item_Identifier', 'string'),
 ('Item_Weight', 'double'),
 ('Item_Fat_Content', 'string'),
 ('Item_Visibility', 'double'),
 ('Item_Type', 'string'),
 ('Item_MRP', 'double'),
 ('Outlet_Identifier', 'string'),
 ('Outlet_Establishment_Year', 'int'),
 ('Outlet_Size', 'string'),
 ('Outlet_Location_Type', 'string'),
 ('Outlet_Type', 'string'),
 ('Item_Outlet_Sales', 'double')]
```

Data Preprocessing

Here we convert the data into machine readable form

VectorAssembler :- It is feature transformer that combine multiple columns into a single vector column.

StringIndexer :- It is use for mapping a string columm to a index column that will be treated as a categorical column by spark.

OneHotEncoder :- It is an important technique for converting categorical attributes into a numeric vector

In [ ]:

```
from pyspark.ml.feature import VectorAssembler, StringIndexer, OneHotEncoder
```

In [ ]:

```
str_index = StringIndexer(inputCols = ['Item_Identifier','Item_Fat_Content','Item_Type',
```

In [ ]:

```
one_hot = OneHotEncoder(inputCols =['Item_Identifier1','Item_Fat_Content1','Item_Type1',
```

In [ ]:

```
vector_ass = VectorAssembler(inputCols = ['Item_Weight','Item_Fat_Content2','Item_Visibi
```

Import Linear Regression and Create Model

In [ ]:

```
from pyspark.ml.regression import LinearRegression
```

In [ ]:

```
linear = LinearRegression(featuresCol="allfeatures", labelCol="Item_Outlet_Sales")
```

Create Pipeline for ML Model

In [ ]:

```
from pyspark.ml import Pipeline
mypipeline = Pipeline(stages = [str_index, one_hot, vector_ass, linear])
```

Making Train Test Split

My Roll No. Is : 37

Using randomsplit data is split into 77% of training and 23% of test as given

In [ ]:

```
training, test = data.randomSplit([0.77, 0.23])
```

Model Training

In [ ]:

```
lin_reg_model = mypipeline.fit(training)
```

Test Model

In [ ]:

```
result = lin_reg_model.transform(test)
```

In [ ]:

```
result.show()
```

```
+---------------+-----------+----------------+---------------+-----------+--------+-----
------------+-------------------------+-----------+--------------------+----------------
-+-----------------+----------------+-----------------+----------+------------------+---
-----------------------+------------+---------------------+------------+----------------
---+-----------------+--------------+------------------+--------------------------+-----
--------+---------------------+-------------+--------------------+------------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|  Item_Type|Item_MRP|Outle
t_Identifier|Outlet_Establishment_Year|Outlet_Size|Outlet_Location_Type|      Outlet_Typ
e|Item_Outlet_Sales|Item_Identifier1|Item_Fat_Content1|Item_Type1|Outlet_Identifier1|Out
let_Establishment_Year1|Outlet_Size1|Outlet_Location_Type1|Outlet_Type1|   Item_Identifi
er2|Item_Fat_Content2|    Item_Type2|Outlet_Identifier2|Outlet_Establishment_Year2| Outl
et_Size2|Outlet_Location_Type2| Outlet_Type2|         allfeatures|        prediction|
+---------------+-----------+----------------+---------------+-----------+--------+-----
------------+-------------------------+-----------+--------------------+----------------
-+-----------------+----------------+-----------------+----------+------------------+---
-----------------------+------------+---------------------+------------+----------------
---+-----------------+--------------+------------------+--------------------------+-----
--------+---------------------+-------------+--------------------+------------------+
|          DRA12|       11.6|         Low Fat|            0.0|Soft Drinks|141.6154|
      OUT045|                     2002|     Medium|              Tier 2|Supermarket Type
1|        3829.0158|          1051.0|              0.0|       8.0|               7.0|
                    7.0|         0.0|                  1.0|         0.0|(1553,[1051],[1.
0])|     (4,[0],[1.0])|(15,[8],[1.0])|     (9,[7],[1.0])|             (8,[7],[1.0])|(2,
[0],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,1,14,21,22...|  2277.372227927723|
|          DRA12|       11.6|         Low Fat|    0.041112694|Soft Drinks|142.0154|
      OUT018|                     2009|     Medium|              Tier 3|Supermarket Type
2|         850.8924|          1051.0|              0.0|       8.0|               5.0|
                    5.0|         0.0|                  0.0|         3.0|(1553,[1051],[1.
0])|     (4,[0],[1.0])|(15,[8],[1.0])|     (9,[5],[1.0])|             (8,[5],[1.0])|(2,
[0],[1.0])|        (2,[0],[1.0])|    (3,[],[])|(29,[0,1,5,14,21,...|   1937.56029983537|
|          DRA12|       11.6|         Low Fat|    0.068535039|Soft Drinks|143.0154|
      OUT010|                     1998|     Medium|              Tier 3|    Grocery Stor
e|         283.6308|          1051.0|              0.0|       8.0|               8.0|
                    8.0|         0.0|                  0.0|         1.0|(1553,[1051],[1.
0])|     (4,[0],[1.0])|(15,[8],[1.0])|     (9,[8],[1.0])|                 (8,[],[])|(2,
[0],[1.0])|        (2,[0],[1.0])|(3,[1],[1.0])|(29,[0,1,5,14,21,...| 327.38010298800816|
|          DRA24|      19.35|         Regular|    0.039920687|Soft Drinks|163.3868|
      OUT035|                     2004|      Small|              Tier 2|Supermarket Type
1|        3439.5228|           322.0|              1.0|       8.0|               1.0|
                    2.0|         1.0|                  1.0|         0.0| (1553,[322],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|     (9,[1],[1.0])|             (8,[2],[1.0])|(2,
[1],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...| 2671.7706482761328|
|          DRA24|      19.35|         Regular|    0.040154087|Soft Drinks|164.6868|
     OUT017|                           2007|     Medium|              Tier 2|Supermarket Type
1|       1146.5076|           322.0|                1.0|       8.0|                 2.0|
                   3.0|         0.0|                     1.0|         0.0| (1553,[322],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|     (9,[2],[1.0])|             (8,[3],[1.0])|(2,
[0],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...|  2691.662378943857|
|          DRA59|       8.27|         Regular|            0.0|Soft Drinks|183.2924|
     OUT017|                           2007|     Medium|              Tier 2|Supermarket Type
1|       2406.2012|            97.0|                1.0|       8.0|                 2.0|
                   3.0|         0.0|                     1.0|         0.0|  (1553,[97],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|     (9,[2],[1.0])|             (8,[3],[1.0])|(2,
[0],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,2,14,21,22...| 3009.0974065421287|
|          DRA59|       12.6|         Regular|    0.127308434|Soft Drinks|186.6924|
     OUT027|                           1985|     Medium|              Tier 3|Supermarket Type
3|       7033.5112|            97.0|                1.0|       8.0|                 4.0|
                   0.0|         0.0|                     0.0|         2.0|  (1553,[97],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|     (9,[4],[1.0])|             (8,[0],[1.0])|(2,
[0],[1.0])|        (2,[0],[1.0])|(3,[2],[1.0])|(29,[0,2,5,14,21,...|    4393.2781578674|
|          DRB01|       7.39|         Low Fat|    0.082367244|Soft Drinks| 187.753|
     OUT049|                           1999|     Medium|              Tier 1|Supermarket Type
1|        1518.024|          1336.0|                0.0|       8.0|                 3.0|
                   4.0|         0.0|                     2.0|         0.0|(1553,[1336],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|     (9,[3],[1.0])|             (8,[4],[1.0])|(2,
[0],[1.0])|            (2,[],[])|(3,[0],[1.0])|(29,[0,1,5,14,21,...|  2996.614390466222|
|          DRB13|      6.115|         Regular|    0.007043008|Soft Drinks| 190.353|
     OUT035|                           2004|      Small|              Tier 2|Supermarket Type
1|         569.259|          1052.0|                1.0|       8.0|                 1.0|
                   2.0|         1.0|                     1.0|         0.0|(1553,[1052],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|     (9,[1],[1.0])|             (8,[2],[1.0])|(2,
[1],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...|  3120.027487849694|
|          DRB13|      6.115|         Regular|     0.01179078|Soft Drinks| 189.053|
     OUT010|                           1998|     Medium|              Tier 3|    Grocery Stor
e|         948.765|          1052.0|                1.0|       8.0|                 8.0|
                   8.0|         0.0|                     0.0|         1.0|(1553,[1052],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|     (9,[8],[1.0])|                 (8,[],[])|(2,
[0],[1.0])|        (2,[0],[1.0])|(3,[1],[1.0])|(29,[0,2,5,14,21,...| 1146.3741600280819|
|          DRB25|       12.3|         Low Fat|    0.069446588|Soft Drinks|106.3938|
     OUT035|                           2004|      Small|              Tier 2|Supermarket Type
1|        857.5504|           323.0|                0.0|       8.0|                 1.0|
                   2.0|         1.0|                     1.0|         0.0| (1553,[323],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|     (9,[1],[1.0])|             (8,[2],[1.0])|(2,
[1],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,1,5,14,21,...| 1711.1967642199536|
|          DRB48|       12.6|         Regular|    0.024733134|Soft Drinks| 40.2822|
     OUT027|                           1985|     Medium|              Tier 3|Supermarket Type
3|       1296.3126|           672.0|                1.0|       8.0|                 4.0|
                   0.0|         0.0|                     0.0|         2.0| (1553,[672],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|     (9,[4],[1.0])|             (8,[0],[1.0])|(2,
[0],[1.0])|        (2,[0],[1.0])|(3,[2],[1.0])|(29,[0,2,5,14,21,...| 2156.0649493401916|
|          DRB48|      16.75|         Regular|    0.024848788|Soft Drinks| 39.9822|
     OUT035|                           2004|      Small|              Tier 2|Supermarket Type
1|        746.3618|           672.0|                1.0|       8.0|                 1.0|
                   2.0|         1.0|                     1.0|         0.0| (1553,[672],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|     (9,[1],[1.0])|             (8,[2],[1.0])|(2,
[1],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...|  769.7994983213209|
|          DRB48|      16.75|         Regular|    0.041599644|Soft Drinks| 40.9822|
     OUT010|                           1998|     Medium|              Tier 3|    Grocery Stor
e|        157.1288|           672.0|                1.0|       8.0|                 8.0|
                   8.0|         0.0|                     0.0|         1.0| (1553,[672],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|     (9,[8],[1.0])|                 (8,[],[])|(2,
[0],[1.0])|        (2,[0],[1.0])|(3,[1],[1.0])|(29,[0,2,5,14,21,...|-1171.6910866866936|
|          DRC01|       5.92|         Regular|    0.019278216|Soft Drinks| 48.2692|
     OUT018|                           2009|     Medium|              Tier 3|Supermarket Type
2|        443.4228|           673.0|                1.0|       8.0|                 5.0|
                   5.0|         0.0|                     0.0|         3.0| (1553,[673],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[5],[1.0])|              (8,[5],[1.0])|(2,
[0],[1.0])|        (2,[0],[1.0])|    (3,[],[])|(29,[0,2,5,14,21,...|  582.7858526373045|
|          DRC01|       5.92|          Regular|     0.019308607|Soft Drinks| 49.0692|
     OUT017|                            2007|      Medium|                 Tier 2|Supermarket Type
1|          1478.076|            673.0|                1.0|       8.0|                 2.0|
                    3.0|            0.0|                 1.0|         0.0| (1553,[673],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[2],[1.0])|              (8,[3],[1.0])|(2,
[0],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,2,5,14,21,...|  929.3680572902323|
|          DRC12|      17.85|          Low Fat|      0.03781972|Soft Drinks|191.6188|
     OUT035|                            2004|       Small|                 Tier 2|Supermarket Type
1|         2475.4444|           1498.0|                0.0|       8.0|                 1.0|
                    2.0|            1.0|                 1.0|         0.0|(1553,[1498],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|      (9,[1],[1.0])|              (8,[2],[1.0])|(2,
[1],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,1,5,14,21,...| 3030.7345765579985|
|          DRC12|      17.85|          Low Fat|     0.037826873|Soft Drinks|189.7188|
     OUT046|                            1997|       Small|                 Tier 1|Supermarket Type
1|         2285.0256|           1498.0|                0.0|       8.0|                 6.0|
                    6.0|            1.0|                 2.0|         0.0|(1553,[1498],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|      (9,[6],[1.0])|              (8,[6],[1.0])|(2,
[1],[1.0])|            (2,[],[])|(3,[0],[1.0])|(29,[0,1,5,14,21,...|  3023.129033573159|
|          DRC12|      17.85|          Low Fat|     0.038040837|Soft Drinks|189.1188|
     OUT017|                            2007|      Medium|                 Tier 2|Supermarket Type
1|         3237.1196|           1498.0|                0.0|       8.0|                 2.0|
                    3.0|            0.0|                 1.0|         0.0|(1553,[1498],[1.
0])|    (4,[0],[1.0])|(15,[8],[1.0])|      (9,[2],[1.0])|              (8,[3],[1.0])|(2,
[0],[1.0])|        (2,[1],[1.0])|(3,[0],[1.0])|(29,[0,1,5,14,21,...| 2991.8000936798494|
|          DRC13|       8.26|          Regular|     0.032573725|Soft Drinks| 125.073|
     OUT018|                            2009|      Medium|                 Tier 3|Supermarket Type
2|           985.384|           1337.0|                1.0|       8.0|                 5.0|
                    5.0|            0.0|                 0.0|         3.0|(1553,[1337],[1.
0])|    (4,[1],[1.0])|(15,[8],[1.0])|      (9,[5],[1.0])|              (8,[5],[1.0])|(2,
[0],[1.0])|        (2,[0],[1.0])|    (3,[],[])|(29,[0,2,5,14,21,...| 1764.2353427575733|
+----------------+-----------+----------------+----------------+-----------+--------+-----
------------+-------------------------+-----------+--------------------+----------------
-+------------------+----------------+------------------+----------+------------------+---
-----------------------+-------------+---------------------+------------+----------------
---+------------------+---------------+-------------------+--------------------------+-----
--------+----------------------+--------------+---------------------+--------------------+
only showing top 20 rows
```

Evaluate Model

In [ ]:

```python
from pyspark.ml.evaluation import RegressionEvaluator
```

In [ ]:

```python
errors = ["r2", "rmse", "mse", "mae"]
name = ["R-Square or Accuracy", "Root Mean Square Error", "Mean Square Error", "Mean Abs

for i in range(len(errors)):
  eval = RegressionEvaluator(predictionCol="prediction", labelCol='Item_Outlet_Sales', m
  print("The {} of Model is {}".format(name[i],eval.evaluate(result)))
```

```
The R-Square or Accuracy of Model is 0.5609324399455548
The Root Mean Square Error of Model is 1146.154277794764
The Mean Square Error of Model is 1313669.6285072374
The Mean Absolute Error of Model is 854.720185337692
```

In [ ]: